CRD

Computational Research Division Report

BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

AUGUST 2007

Genome Matchmaker

Harvard grad student devises an efficient method to sort and organize billions of genomic matches

A DOE graduate fellow has developed an algorithm that will dramatically slash the time it takes to sort and catalog billions of genome sequences from the Joint Genome Institute and other research centers.

The algorithm, developed by Ben Campbell Smith from Harvard University, can search and organize billions of genomic sequence comparisons in a day instead of a month. The efficiency will enable staff at the Biological Data Management and Technology Center (BDMTC) at Berkeley Lab to massage raw data into materials that scientists can easily use for genomic analyses.

continued on page 2

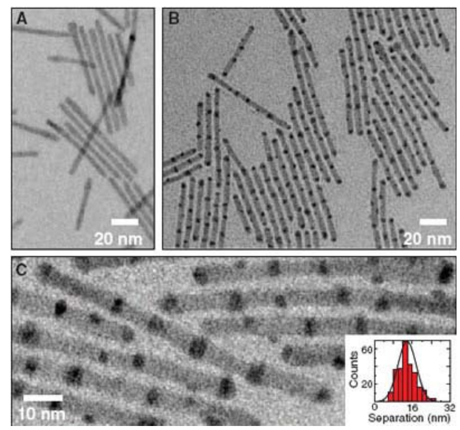
Connecting Dots

Computational research leads to new method for creating striped nanorods

Using methods developed by CRD scientists Denis Demchenko and Lin-Wang Wang, a research team has found a way to make striped nanorods in a colloid, a suspension of particles in solution, according to a paper published in the journal Science last month.

The contributions from Demchenko and Wang enabled their fellow research team members from Berkeley Lab and UC Berkeley to carry out the nanorod project and performed calculations on supercomputers at the National Energy Research Scientific Computing Center.

"This project has involved tight coordination between computer simulations and experiment, and the results obtained here would not have been possible to achieve without the contributions of our computational scientists, Denis Demchenko and Lin-Wang Wang," said Paul Alivisatos, lead researcher and the director of the Materials Sciences Division at Berkeley Lab. "It is another clear example where we see that theoretical simulations are not just being used to explain materials



In these Transmission Electron Microscope images of superlatticed or striped nanorods formed through partial cation exchange, (A) shows the original cadmium-sulfide nanorods; (B and C) show cadmium-sulfide nanorods striped with silver-sulfide. The inset is a histogram showing the pattern spacing of the silver-sulfide stripes.

growth after the fact, but are now an integral part of the materials design and creation process from the very start."

Superlatticed or "striped" nanorods – crystalline materials only a few molecules in thickness and made up of two or more semiconductors – are highly valued for their potential to serve in a variety of devices, including transistors, biochemical sensors and light-emitting diodes (LEDs).

Until now the potential of superlatticed nanorods has been limited by the relatively expensive and exacting process required to make them. Previously, striped nanorods were made through epitaxial processes, in which the rods were attached to or embedded within a solid medium.

One of the key differences between quantum dots epitaxially grown on a substrate and free-standing colloidal quan-

continued on page 3

Man Who Loves Numbers

New computational science fellow aims to broaden his research in chemistry, nanoscience and beyond

It was his love for math and engineering – and a determination to pursue a better education – that prompted George Pau to leave Malaysia four years ago for the United States. That same drive for promising opportunities brought him to Berkeley Lab a month ago as the new Luis W. Alvarez Fellow in Computational Science.

Pau, 29, came here after earning his Ph.D. in mechanical engineering at the Massachusetts Institute of Technology, where he was drawn into the world of developing numerical methods for solving chemistry problems. As a fellow, Pau plans to broaden his research scope to include the understanding of multiscale phenomena in physical systems and the exploration of how numerical methods could enhance nanostructure designs.

continued on page 4

Genome continued from page 1

BDMTC develops informatics tools and provides data management for the Joint Genome Institute (JGI), UC San Francisco, Berkeley Lab's Life Sciences and Physical Biosciences divisions and the California Institute of Quantitative Biomedical Research (QB3). The Integrated Microbial Genomes (IMG) system, created by BDMTC, integrates microbial data from the JGI and other public sources and enables comparative analyses across species, something researchers look for in hunting for clues about evolution, for example.

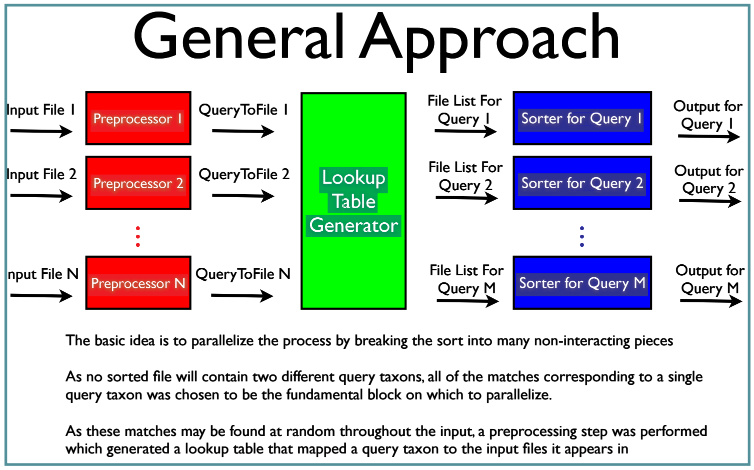
"Ben is an outstanding worker. Bioinformatics is now flooded with a huge amount of data to be analyzed and cross compared. Scalability is a major issue especially for a tightly integrated system such as IMG," said Ernest Szeto, a BDMTC researcher who works closely with Smith. "Ben has applied solid computer science skills to help deal with some of these most pressing disk oriented processing scalability issues."

Smith is working in BDMTC this summer as a DOE Computational Science Graduate Fellow. The fellowship, funded by the DOE Office of Science and the National Nuclear Security Administration, not only pays for each fellow's tuition and other school fees, it also provides an annual stipend of $31,200 and other funds for research-related expenses.

Bioinformatics is not Smith's research focus in school. In fact, the Harvard graduate student is partial to high-energy physics. His Ph.D. work involves hunting for the elusive Higgs boson particle, whose existence can validate a theory on how fundamental particles such as electrons and quarks acquire mass.

But working with biological data isn't new for Smith. He recalled fondly the time he worked in his father's bioinformatics lab at the University of British Columbia, where the elder Smith is a hematologist/oncologist.

"Before I started graduate school, my dad said, 'Come work for me and write some code.' I had a lot of fun doing that," said Smith, who searched Berkeley Lab's Computing Sciences web site for research ideas and learned about the work by Markowitz and his group. "I thought it would be cool to work on something that people use all the time."



Ben Campbell Smith's algorithm first breaks down the large text file containing billions of genomic matches before it begins the sorting process.

Genomics and high-energy physics share one similarity – they both generate an incredible amount of research data that must be culled to obtain useful information for research. With that in mind, Smith said he was able to immerse himself quickly in the informatics project at Berkeley Lab.

Every time a microbe's genome is sequenced, that information goes to the Pacific Northwest National Laboratory (PNNL), which uses a supercomputer and a software called Basic Local Alignment Search Tool (BLAST) to look for matching sequences among the roughly 3.2 million microbial sequences in the database.

Instead of looking for matches only between the newly sequenced genome and those already in the database, however, the PNNL computer carries out the "all versus all" BLAST search, spitting out results that show all the matches among various microbes' genomes. When each microbe's genome can produce thousands of gene sequences, the process of matching them with each other will produce an enormous set of data. As a result, BDMTC staff aren't able to update the IMG system frequently.

Smith's task is to organize and format those results so that researchers can quickly find specific comparisons among the two sequences or microbes they are studying. The dataset he is working with contains 20 billion lines, each corresponding with a match.

The 20 billion matches aren't in any particular order, making it even more difficult to sort them by taxons and then "score," which refers to a statistical analysis of the quality of the matches (some matches could have been made in error).

Before Smith devised the new method, BDMTC staff used a brute force algorithm that read the output a single line at a time and wrote the match to a file based upon the two taxons involved. Because of the inefficiency inherent in accessing a different file for each of the nearly 20 billion matches, this process would take approximately 30 days.

Smith's algorithm, on the other hand, first breaks down the data into thousands of smaller files. Using a cluster with 35 dual core CPUs, the smaller chunks of data are catalogued by the genomes they contain. This allows a sorting program to focus only on very small subsets of the data corresponding to the genome of interest. The process is further sped up through the use of a binary search tree, which allows the sorting to remain computationally efficient, even for very large datasets.

"The process now takes a day. You take all the data and run and sort it. Then anyone who needs it again can quickly look up the results," Smith said.

With the new technique, the IGM system can be updated four times a year instead of two. The algorithm will be used in the next release of IMG/M, the metagenomics version of IMG, which is accompanied with a big batch of computational results from PNNL. The next release is scheduled for December or January.

Learn more about the IMG system at http://crd.lbl.gov/html/BDMTC. Information about the Computational Science Graduate Fellowship program can be found at http://www.krellinst.org/csgf/index.shtml.

Nanorod *continued from page 1*

tum dots is the presence of strain. The use of temperature, pressure and other forms of stress to place a strain on material structures that can alter certain properties is called “strain engineering.” This technique is used to enhance the performance of today’s electronic devices, and has recently been used to spatially pattern epitaxially grown striped nanorods.

However, strain engineering in epitaxially produced striped nanorods requires clever tricks, whereas Demchenko, Wang and their colleagues discovered – through ab initio calculations of the interfacial energy and computer modeling of strain energies – that naturally occurring strain in the colloidal process would be the driving force that induced the spontaneous formation of the superlattice structures.

“We have studied structure and elastic properties of these superlattices, as well as their electronic structure. In particular, the machanism of formation and ordering of these nanorod superlattices, initially puzzling, was explained by using large nanoscale elastic properties calculations,” said Demchenko, a member of the Scientific Computing Group in CRD. “It turned out that large strain fields created by the lattice mismatch at the CdS/Ag_2S interface are responsible for the formation of the ordered nanorod superlattice. This is an interesting result because it has not been previously observed in quasi-1D geometry of the nanorod.”

The paper in Science, titled “Spontaneous Superlattice Formation in Nanorods Through Partial Cation Exchange,” also was co-authored by Richard Robinson of Berkeley Lab’s Materials Sciences Division, as well as Bryce Sadtler and Can Erdonmez of UC Berkeley’s Department of Chemistry.

Today’s electronics industry is built on two-dimensional semiconductor materials that feature carefully controlled doping and interfaces. Tomorrow’s industry will be built upon one-dimensional materials, in which controlled doping and interfaces are achieved through superlatticed structures. Formed from alternating layers of semiconductor materials with wide and narrow band gaps, superlatticed structures, such as striped nanorods, not only can display outstanding electronic properties, but photonic properties as well.

Previous research by Alivisatos and his group had shown that the exchange of cations could be used to vary the proportion of two semiconductors within a single nanocrystal without changing the crystal’s size and shape, so long as the crystal’s minimum dimension exceeded four nanometers. This led the group to investigate the possibility of using a partial exchange of cations between two semiconductors in a colloid to form a superlattice. Working with previously formed cadmium-sulfide nanorods, they engineered a cation exchange with free-standing quantum dots of the semiconductor silver-sulfide.

Even though the colloidal striped nanorods form spontaneously, Alivisatos said it should be possible to control their superlatticed pattern – hence their properties – by adjusting the length, width, composition, etc., of the original nanocrystals. However, much more work remains to be done before the colloidal method of fabricating striped nanorods can match some of the “spectacular results” that have been obtained from epitaxial fabrication.

More information about the research is at http://www.cchem.berkeley.edu/~pagrp.

Software Tools for Code Improvement

For the eighth consecutive year, Tony Drummond and Osni Marques of CRD’s Scientific Computing Group hosted a workshop, which took place this month, that focused on the DOE Advanced CompuTational Software (ACTS) Collection. ACTS comprises a set of non-commercial tools that enable researchers to improve their codes.

The four-day workshop, entitled “Building Blocks for Reliable and High Performing Computing,” drew nearly 40 people. The attendees listened to presentations each morning and received hands-on training using NERSC supercomputers in the afternoon. The ACTS Collection is developed mainly at DOE laboratories, with collaboration with universities.

The workshop presented an introduction to the ACTS Collection for scientists and graduate students whose research require them to deal with large amounts of computation, complex software integration, distributed computing and robust numerical algorithms. The workshop included a range of tutorials and tools developed by the DOE SciDAC Program, as well as discussion sessions aimed to solve specific computational problems.

The ACTS software tools simplify the solution of common and important computational problems. They not only enable applications to run efficiently on high performing computing environments, but they also enable computation that would not have been possible otherwise. Learn more about the workshop at http://acts.nersc.gov/events/Workshop2007.

Berkeley Lab researchers Osni Marques (right, middle photo) and Tony Drummond (left, middle photo) help participants in a software training workshop.

Stellar Newcomer

New postdoc brings fresh approach to computational research on Type Ia supernovae

Andy Nonaka has joined the Center for Computational Sciences and Engineering (CCSE) as a postdoc, focusing his research on low Mach number hydrodynamics codes for stellar environments, particularly those found in Type Ia supernovae. Nonaka, who recently earned a Ph.D. in Engineering Applied Science at UC Davis, joined Berkeley Lab last month and is working closely with scientists John Bell and Ann Almgren in CCSE.

Nonaka, 27, is no stranger to Berkeley Lab. He was a guest researcher working with Phil Colella in the Applied Numerical Algorithms Group, also part of CRD, while working on his dissertation, "A Higher-Order Upwind Method for Viscoelastic Fluids." The work centered on exploring the fluid dynamics of non-Newtonian fluids in irregular geometries, particularly those found in bioMEMS devices.

The transition from his Ph.D. work to his current research has been very smooth, he said.

"The physics are different, but the methods you use to solve both problems are very similar," said Nonaka, who credits his thesis advisor at UC Davis, Greg Miller, for introducing him to computational science.

He was awarded a Student Employee Graduate Research Fellowship from Lawrence Livermore National Laboratory that provided four years of funding for his graduate research. At Livermore, he worked with David Trebotich in the Institute of Scientific Computing Research.

Nonaka also worked as a laser diagnostic technician during an internship at Livermore Lab when he was an undergraduate student in the Electrical Engineering program at the University of the Pacific in Stockton, about 75 miles east of Berkeley. As a diagnostic technician, Nonaka worked on the design, construction, operation and repair of laser imaging instruments, including CCD cameras.

When he's not working at the Lab, you may find him perfecting his strokes at the pool. Nonaka, a champion swimmer, competed as an NCAA Division I athlete as a member of the Men's Intercollegiate Swim Team at Pacific. The sport brought good luck for Nonaka – he met his wife and fellow swimmer, Marissa, on the Davis Aquatic Masters Adult Swim Team while attending UC Davis. The couple is expecting their first child in February.

Andy Nonaka

Now Nonaka is a member of the Walnut Creek Masters, which is part of Pacific Masters Swimming, a local governing body that oversees more than 10,000 members in Northern California and Nevada. He holds the Pacific Masters record for the 1,500-meter freestyle for the age 25-29 group. He also is a member of the teams holding three national relay records for the age 19-24 group and has been named Swimmer of the Year for his age group in each of the past 6 years.

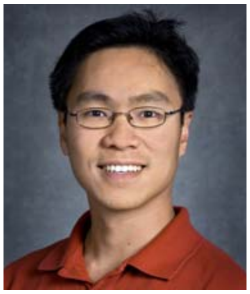
George Pau *continued from page 1*

"I am glad to be here at LBL because scientists here collaborate extensively on developing codes. It allows me to see how it can be done," said Pau.

When he applied for the fellowship, he wanted to work with John Bell, director of the Center for Computational Sciences and Engineering in the Computational Research Division. Pau got his wish and now works out of a wing in Building 50A, among researchers who develop algorithms and mathematical models for combustion, supernovae and subsurface flow, for example.

"George is a pleasure to work with. Although he has only been at LBNL for a month, he is already making significant contributions to our research programs," said Bell.

Polite and engaging, Pau grew up in Malaysia, in the city of Miri on the northwestern shore of the Borneo island. Being a doctor or an engineer confers a prestige not available to other professions in Malaysia. So when it was time to decide on a major in college, Pau picked mechanical engineering. It wasn't a tough choice.



George Pau

"I can't become a doctor because I don't like the sight of blood, so the only choice was engineering," Pau said with a smile.

He excelled in school and won a scholarship and a promise from the government to send him to the United Kingdom for undergraduate study. But the government found itself lacking money to fulfill that promise when the 1997 Asian economic crisis hit. So it placed Pau in a newly built university. Pau was part of the first graduating class at the Petronas University of Technology in 2001, when he earned a degree in mechanical engineering.

What to do next? Malaysia's higher-education system couldn't offer him rigorous course work beyond college, he said. "I've always wanted to get a Ph.D., so I decided to go to Singapore for a master's degree," Pau said.

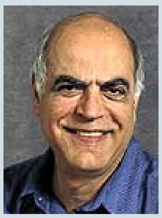
The neighboring country gave him a

continued on page 6

Hall of Fame

Software to Streamline Data Management

A paper co-authored by Arie Shoshani on "Storage Resource Managers: Recent International Experience on Requirements and Multiple Co-Operating Implementations" has been accepted for publication at the 24th IEEE Conference on Mass Storage Systems and Technologies in San Diego next month.



Arie Shoshani

Shoshani, head of the Scientific Data Management Research Group in CRD, also is the coordinator and editor of this paper, which addresses the challenges of dealing with a variety of storage systems during a large collaborative research. This paper summarizes the implementation of software developed for multiple, heterogeneous storage systems using a common interface standard called Storage Resource Manager (SRM).

The publication also describes using SRM in a large high-energy physics collaboration among eight institutions in the United States and Europe. The project, called WLCG, set out to prepare for the handling of large volumes of data that will be generated when the Large Hadron Collider (LHC) goes online at CERN near Geneva next year.

Shoshani and his co-authors also talk about the setup and running of a large number of compatibility tests, which are carried out several times a day. Several of the scientists in Shoshani's Data Management Research Group also contributed to the paper: Alex Sim, Vijaya Natarajan and Junmin Gu. More information about the conference is at http://storageconference.org/2007.

A Fellow in Japan

The Japan Society for the Promotion of Science has awarded a fellowship to Osni Marques. The fellowship allows Marques, a researcher in the Scientific Computing Group in CRD, to spend 60 days at a research institution in Japan. Marques plans to do so in February and March next year and will be based at the University of Tokyo.

The fellowship program, called "Invitation Fellowship for Research in Japan," encourages Japanese scientists to invite their foreign colleagues and work more closely in research and academic activities.

"It's a good opportunity for me to pursue collaborations in Japan, and I am looking forward to it," Marques said.

More information about the fellowship and other programs offered by the society can be found at http://www.jsps.go.jp/english.

Indexing Technology Speeds Up Data Retrieval

CRD scientists published a paper on data management research at the 19th International Conference on Scientific and Statistical Database Management, which took place in Banff, Canada last month.

The paper, "Enabling Real-Time Querying of Live and Historical Stream Data," was co-authored by Arie Shoshani, Frederick Reiss, Kurt Stockinger and Kesheng Wu from Berkeley Lab. A fifth author, Joseph M. Hellerstein, is a researcher at UC Berkeley's Computer Science Division.

The publication described the researchers' bitmap indexing technology called FastBit, a more efficient method of updating archival data. Right now, software programs that query data streams in order to identify trends, patterns or anomalies could offer more values if they are able to compare the live stream data with historical data in archives.

But searching for the historical data in real time is prohibitively expensive, partly because of the cost of updating the indices for the archival data. FastBit promises to solve that bottleneck.

Find out more about data management research at http://ssdbm2007.cpsc.ucalgary.ca.

Chris Ding Leaves to Teach

Chris Ding

Chris Ding, one of the founding members of the Scientific Computing Group when NERSC was relocated to the Lawrence Berkeley National Laboratory in 1996, has accepted a faculty position at the University of Texas at Arlington (UTA). He joins the Department of Computer Science and Engineering at UTA as a tenured full professor this August.

Over his 11 years at Berkeley Lab, Ding has been a key contributor and intellectual leader in the Scientific Computing Group, said Esmond Ng, leader of the group. His research portfolio at the lab has included high performance computing, climate simulation, data mining/analysis and bioinformatics. "In particular, Chris was instrumental in bringing the SciDAC climate project to LBNL in 2001," Ng said. "He has also been successful in establishing research collaborations with several divisions at LBNL, such as the Earth Sciences Division and the Physical Biosciences Division."

Ding is expecting to continue his collaborations with several staff members at Berkeley Lab after he moves to UTA. Because of that, he will be

continued on page 6

Hall of Fame

continued from page 5

appointed as a participating guest in the Scientific Computing Group so that he can visit and collaborate with scientists here as often as his time permits.

Mathematicians Meet in Zurich

The 6th International Congress on Industrial and Applied Mathematics in Zurich last month featured several CRD researchers as speakers.

Xiaoye Li and Osni Marques organized a minisymposium on "Mathematical Algorithms, Frameworks and Scientific Applications on Large-Scale Parallel Machines." CRD scientists who presented papers during the minisymposium included John Bell, Juan Meza and Andrew Canning.

John Bell

Bell, Director of the Center for Computational Sciences and Engineering, also spoke in three other minisymposia. His presentations were:

* Low Mach-number models in computational astrophysics.

* Modeling of fluctuation in algorithm refinement methods.

* Simulation of lean pre-mixed turbulent combustion.

The conference attracted more than 3,000 registrants from the Americas, Europe, Asia, Africa and Australia. Learn all about the hot topics in the world of applied math by clicking http://www.iciam07.ch/index.

Bro Workshop Drew Sell-Out Crowd

Berkeley Lab researchers Brian Tierney, Vern Paxson, Robin Sommer and Scott Campbell held a three-day workshop on the Bro Intrusion Detection System last month in San Diego, which attracted more participants than anticipated.

The workshop, featuring tutorials on Bro installation and customization, drew 31 attendees from not only the United States but also Europe and South America. Organizers had planned the workshop to accommodate 30 people.

"We got very positive feedback on the workshop, and will likely do another one at LBNL in 6-12 months," said Tierney, head of the Collaborative Computing Technologies Group in the Computational Research Division.

Developed by Paxson and other scientists at Berkeley Lab, Bro is an open-source, UNIX-based system that passively monitors network traffic and looks for suspicious activities. The Lab first deployed Bro in 1996, and it has since become the cornerstone of LBNL's cyber security defenses. Learn about Bro at http://www.bro-ids.org.

The workshop, which took place at the San Diego Supercomputer Center, drew eight participants from DOE and NASA labs, four from industry and the remainder from universities.

The workshop agenda, with links to presentations, can be found at http://www.bro-ids.org/bro-workshop-2007/agenda.htm.

George Pau *continued from page 4*

great opportunity – Pau enrolled in the Singapore-MIT Alliance program at the National University of Singapore. The alliance started in 1998 to provide educational and research collaboration in engineering and life science.

It was at the National University where Pau first learned about numerical methods, which enable scientists to understand the properties and behavior of a phenomenon through computer simulations. This research area has broad applications in many industries, including the discovery of stable chemical compounds and improved turbine designs in power plants.

MIT accepted Pau into its Ph.D. program after he worked for Motorola in Singapore for a year. At MIT, Pau became interested not only in developing numerical methods but also in analyzing existing methods and improving their efficiency. Quantifying and reducing calculation errors are some of the challenging problems for Pau to tackle during his fellowship.

Moving across the country for the fellowship is just one of the big changes in his life. Two months ago, Pau married Ming Lee Tang, a fellow Malaysian who is now a graduate student in chemistry at Stanford University. The couple met when Tang attended Brandeis University, near MIT.

Outside of work, Pau enjoys hiking, sea kayaking and tennis. He also loves to cook and counts beef rendang, along with braised duck in a five-spice powder and galanga (a root that resembles ginger) concoction, as some of his signature dishes.

About CRD Report

CRD Report, which publishes every other month, highlights the cutting-edge research conducted by staff scientists in areas including turbulent combustion, nano materials, climate change, distributed computing, high-speed networks, astrophysics, biological data management and visualization. CRD Report Editor Ucilia Wang can be reached at 510 495-2402 or Uwang@lbl.gov.
Find previous CRD Report articles at http://crd.lbl.gov/html/news/CRDreport.html.